Role of Morpho-Syntactic Features in Estonian Proficiency Classification

Sowmya Vajjala and Kaidi Lõo
Department of Linguistics, University of Tübingen

Overview

- We developed an approach to perform proficiency classification for learners of Estonian as a second language.
- Using a publicly accessible Estonian learner corpus, we show that
  - morpho-syntactic features in learner texts are useful predictors.
  - cascades of binary classifiers perform better than performing the classification in a single step.

Related Work

- SLA researchers studied the characteristic features of learner texts at different proficiency levels. (e.g., Tono, 2000; Vyatkina, 2012; Lu, 2012)
- Automated assessment of student essays is also an active research area. (e.g., Yannakoudakis, Briscoe & Medlock, 2011; Burstein, 2013)
- Contemporary research primarily focused on learner errors across proficiency levels. (e.g., Dickinson, Kübler & Meyer, 2012)
- But, the role of morpho-syntactic features in proficiency classification was not explored before.

Estonian Morphology

- Estonian is agglutinative. Word forms can be formed by joining the morphemes together.
  - e.g., *jalgades –>jalga+de+s* (stem for foot +plural marker+inessive case marker)
- It is fusional i.e., word forms can be formed by changing the stem.
  - e.g., *jalg* (foot, nominative), *jala* (genitive), *jalga* (partitive)
- It has 14 productive cases (grammatical and semantic cases).
  - Cases express relations between words and are sometimes used instead of postpositions (*jalal* and *jala peal* have the same meaning: *on the foot*)
- Cases have different alternative case endings.
  - e.g., Valid allative plural forms for *jalg* (foot) are: *jalgadele, jalule, jalgele*

- We model some of these morphological characteristics as features for the learner proficiency classification task.

The Corpus

- The Estonian Interlanguage Corpus (EIC) consists of texts written by learners of Estonian as a Second Language (Eslon, 2007).
- It mainly consists of short answers, essays and personal letters.
- It also has error annotations but we did not use them in this paper.
- Here is a numeric description of the corpus:

Proficiency Level	*# Docs*	*Avg. tokens per doc.*
A	807	182.9
B	876	260.3
C	307	431.8

- We created a randomly picked held-out test set with 50 documents per class from this dataset.

Features

Morphological Features

- Nominal inflection features: proportion of nouns and adjectives tagged with various cases.
- Verbal inflection features: proportion of verbs belonging to various tense, mood, voice, number and person categories.

Other Features

- POS features: proportion of words of various parts of speech
- Lexical variation features: ratio of nouns, verbs, adjectives and adverbs to lexical words (Lu, 2012)
- Text length: number of word tokens per text

BEST10FEATURES were determined automatically.

- selection method: Correlation based Feature Subset (CFS) selection
- ranking method: Information Gain

Feature	**Group**
Nominative case	NounMorph
Impersonal Voice	VerbMorph
Personal Voice	VerbMorph
Num. words	**TextLength**
Present tense	VerbMorph
2nd person verbs	VerbMorph
Prepositions	**POS**
Allative case	NounMorph
Imperatives	VerbMorph
Translative case	NounMorph

Eight of the ten best features are from morphological features group.

Experimental setup

- We approached three class classification using
  - a single classifier (SMO) - with various feature combinations.
  - a Stacking ensemble with SMO, Logistic Regression and Random Forest classifiers (with all features).
  - two class cascade combinations (SMO - with all features) : since binary classification was more accurate.
    - Cascade-1: using the classifiers AC, AB and BC.
      1. Classify the instance using the classifier (A,C).
      2. If A, re-classify using (A,B). Else, re-classify using (B,C).
    - Cascade-2: using the classifiers A-NotA, B-NotB, C-NotC.
      1. Classify the instance using the classifier (C,NotC).
      2. If NotC, re-classify using (A,NotA).
    - The choice of these cascades was primarily heuristic.
- Evaluation Metric: classification accuracy (with both CV and test set)
- All the classifiers had equal number of documents belonging to the classes they are made of.
- The held-out test set was not used in any training stage.

Results

- Binary classification

Classifer	**Accuracy** (10 fold CV)
A vs B	70.8%
B vs C	74.59%
A vs C	85.93%
A vs NotA	74.20%
B vs NotB	60.04%
C vs NotC	79.69%

- Three class classification - a comparison of features and approaches

Classifer	**Accuracy on Test Set**
With All Features	59.33%
Noun+Verb Morph. Features	58%
Best 10 Features	56.66%
Ensemble classifier	57.33%
Cascade Classifier 1	**64.66%**
Cascade Classifier 2	**66.66%**

- Experimenting with different training data sizes showed that it did not have a major impact on classification accuracy.

Conclusions

- Morphological complexity based features indeed play an important role in Estonian proficiency classification.
- Reformulating the three-class classification problem as a cascade of binary classifiers improved the classification accuracy.
- Increasing the training data did not improve the classification accuracy. So, the morphological features are good but not self-sufficient.
- The accuracies we achieved (60-65%) are a good starting point in moving towards a real word application.

Future Work

- Explore other classes of features for this task. e.g., syntactic complexity, error rate, coherence etc.
- Apply insights from SLA research in proficiency classification.
- Explore cascade models better in this context.

References

Burstein, J. (2013). Automated Essay Evaluation and Scoring. In *The Encyclopedia of Applied Linguistics*, Blackwell Publishing Ltd.

Dickinson, M., S. Kübler & A. Meyer (2012). Predicting Learner Levels for Online Exercises of Hebrew. In *Proceedings of the Seventh BEA Workshop*. pp. 95–104.

Eslon, P. (2007). Õppijakeelekorpused ja keeleõpe. In *Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine*. pp. 87–120.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Languages Journal* .

Tono, Y. (2000). A corpus-based analysis of interlanguage development: analysing POS tag sequences of EFL learner corpora. In *PALC'99: Practical Applications in Language Corpora*. pp. 323–340.

Vyatkina, N. (2012). The development of second language writing complexity in groups and individuals: A longitudinal learner corpus study. *The Modern Language Journal* .

Yannakoudakis, H., T. Briscoe & B. Medlock (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of ACL-HLT*. pp. 180–189.